A Survival-Adjusted Quantal-Response Test for Analysis of Tumor Incidence Rates in Animal Carcinogenicity Studies

Shyamal D. Peddada and Grace E. Kissling

Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA

In rodent cancer bioassays, groups of animals are exposed to different doses of a chemical of interest and followed for tumor occurrence. The resulting tumor rates are commonly analyzed using a survival-adjusted Cochran-Armitage (CA) trend test. The CA trend test has reasonable power when the tumor-response curve is linear in dose, but it may be underpowered for a nonlinear response. An alternative survival-adjusted test procedure based on isotonic regression methodology has previously been proposed. Although this alternative procedure performs well when the tumor response is nonlinear in dose, it has less power than the CA trend test when the response is linear in dose. Here, we introduce a new survival-adjusted test procedure that makes use of both the CA trend test and the isotonic regression-based trend test. Using a broad range of experimental conditions typical of National Toxicology Program (NTP) bioassays, we conducted extensive computer simulations to compare the false-positive error rate and power of the proposed procedure with the survival-adjusted CA trend test. The new procedure competes well with the survival-adjusted CA trend test when observed tumor rates are linear in dose and performs substantially better when observed tumor rates are nonlinear in dose. Further, the proposed trend test almost always has a smaller false-positive rate than does the survival-adjusted CA trend test. We also developed an order-restricted inference-based procedure for performing multiple pairwise comparisons between each of the dose groups and the control group. The trend test and the multiple pairwise comparisons test are demonstrated using an example from a study conducted by the NTP. ***Key words:*** **cancer bioassay, Cochran-Armitage trend test, multiple pairwise comparisons, National Toxicology Program, order-restricted inference, poly-3 trend test.** ***Environ Health Perspect*** **114:537–541 (2006). doi:10.1289/ehp.8590 available via** ***http://dx.doi.org/*** **[Online 10 November 2005]**

A major responsibility of the National Toxicology Program (NTP) is to investigate the potential toxic and carcinogenic effects of various chemicals. Studies conducted by the NTP are used by the U.S. Environmental Protection Agency, the Food and Drug Administration, and other federal and state agencies in their consideration of regulations and policies for protecting public health. The 2-year rodent cancer bioassay is an important component of the NTP's investigations; these bioassays typically involve groups of male and female mice and rats randomly assigned to either a control group or one of three dose groups (low, medium, and high). At death, each animal is extensively examined for microscopic and macroscopic tumors, as well as for abnormal changes in tissues. Of particular interest is whether the rate of occurrence of a specific tumor is related to dose.

Because some animals do not survive to the end of the 2-year study, the NTP's statistical analyses of site-specific tumor rates employ a survival-adjusted, continuity-corrected Cochran-Armitage (CA) trend test, the poly-3 trend test (Armitage 1955; Bailer and Portier 1988; Cochran 1954; Portier and Bailer 1989). For simplicity, in this article we refer to this test as the NTP trend test. Typically, this trend test is followed by pairwise comparisons of each dose group with the control group.

It is important to distinguish between two types of parameters related to the problem of current interest, namely, age-specific tumor incidence rate and lifetime tumor rate. The former is the hazard rate associated with the age at tumor onset, and the latter is the expected proportion of animals that ever develop a tumor (at any age), which is a function of both the tumor incidence rate and the mortality rate. Thus, depending on mortality patterns, lifetime rates generally will not correspond exactly with age-specific incidence rates. The age-specific incidence rate is a more meaningful, but less accessible, end point than is the lifetime rate. For simplicity of notation, we have dropped "age-specific" from the phrase "age-specific tumor incidence rate."

All point estimators discussed in this article and in other reports, such as Bailer and Portier (1988) and Peddada et al. (2005), estimate lifetime tumor rates. Further, trend tests such as the NTP trend test, the trend test developed by Peddada et al. (2005), and the trend test proposed in this article use survival-adjusted lifetime tumor rates to test for trends in tumor incidence rate. Simulation studies performed by Dinse (1991) and Peddada et al. (2005) and in this article suggest that, although these tests employ lifetime tumor rates rather than tumor incidence rates, they also provide valid inferences about tumor incidence rates.

Although tumor incidence rates may strictly increase with dose, lifetime tumor rates may have an umbrella-shaped or a plateau-shaped dose–response curve because of higher mortality in the upper dose groups. An umbrella or plateau-shaped response curve for lifetime tumor rate is also likely to occur if the tumor incidence rate has a plateau-shaped dose–response curve. In such situations, the NTP trend test may not be sensitive enough to detect this dose-related response because it is based on linear regression of the estimated lifetime tumor rate on dose. For example, in a study of the chemical isoprene, female rats were exposed to 0, 220, 700, or 7,000 ppm isoprene by inhalation for 2 years (NTP 1999). Mammary gland fibroadenomas occurred in 19, 35, 32, and 32 of 50 animals, respectively. Survival did not differ among dose groups, and the survival-adjusted lifetime tumor rates were 44%, 74%, 74%, and 73%, respectively. The NTP trend test yielded a p-value of 0.11. However, each of the pairwise comparisons with the control group was significant at $p < 0.002$ (NTP 1999). Data such as these are not uncommon in NTP studies, and in these situations the NTP trend test may fail to detect a significant chemical effect.

Motivated by such examples, Peddada et al. (2005) developed an order-restricted inference-based procedure that is well suited for nonlinear responses. Simulation studies suggest that this isotonic regression-based test performs better than does the NTP trend test when the tumor rates increase nonlinearly in dose (Peddada et al. 2005). However, this test may have less power than the NTP trend test when the tumor response increases linearly with dose. Therefore, in this article we propose a new test that modifies their isotonic regression-based

Address correspondence to S.D. Peddada, Biostatistics Branch, MD A3-03, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709 USA. Telephone: (919) 541-1122. Fax: (919) 541-4311. E-mail: peddada@niehs.nih.gov

We thank G. Dinse and B. Gladen for their helpful comments. We also thank the reviewers and the editor for several useful comments that have improved the article.

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences, National Institutes of Health.

The authors declare they have no competing financial interests.

Received 16 August 2005; accepted 9 November 2005.

test. Based on our extensive simulations, the resulting test competes well with the NTP trend test for strictly linear dose–response patterns and performs better than the NTP trend test for nonlinear dose–response patterns.

In addition to testing for dose-related trends, NTP performs pairwise comparisons between each of the dose groups and the control group, with particular attention paid to the medium- and high-dose groups. Currently, no adjustment is made for multiple comparisons. Consequently, the NTP's pairwise comparison tests are subject to false-positive rates that exceed the nominal 0.05 level. Using the order-restricted inference methodology developed in Hwang and Peddada (1994) and Peddada et al. (2001), we introduce a new pairwise comparison procedure that controls the overall false-positive rate.

Materials and Methods

Suppose there are K dose groups with doses $0 = d_1 < d_2 < . . . < d_K$ with n_i animals assigned to the ith dose group, $i = 1, 2, . . . , K$. We denote the tumor status at necropsy of the jth animal in the ith dose group by y_{ij}, where $y_{ij} = 1$ if the animal has a specific tumor and is 0 otherwise. To compute the poly-3 survival-adjusted sample size for the ith group (Bailer and Portier 1988; Portier and Bailer 1989), we define animal-specific weights $w_{ij} = 1$ if the jth animal had a tumor at necropsy, otherwise $w_{ij} = t_{ij}^3$ where t_{ij} is the fraction of duration of the study for which the animal survived. For the ith dose group, the poly-3 survival-adjusted sample size is

$$n_i^* = \sum_{j=1}^{n_i} w_{ij},$$

and the poly-3 survival-adjusted estimator for the lifetime tumor rate, π_i, is

$$\hat{\pi}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i^*.$$

Test for an increasing trend in dose. The poly-3 survival-adjusted CA trend test statistic is

$$T_0 = \frac{\sum_{i=1}^{K} a_i d_i \hat{\pi}_i - \left[\sum_{i=1}^{K} a_i d_i \sum_{i=1}^{K} a_i \hat{\pi}_i\right] \Big/ \left[\sum_{i=1}^{K} a_i\right]}{S\sqrt{\left\{\sum_{i=1}^{K} a_i d_i^2 - \left[\sum_{i=1}^{K} a_i d_i\right]^2 \Big/ \left[\sum_{i=1}^{K} a_i\right]\right\}}}$$

$$= \frac{A}{B}, \qquad [1]$$

where, for $i = 1, 2, . . . , K$,

$$\hat{\pi} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij} \Big/ \sum_{i=1}^{K} n_i^*,$$

$$r_{ij} = y_{ij} - \hat{\pi} w_{ij},$$

$$\overline{r_i} = \sum_{j=1}^{n_i} r_{ij} \Big/ n_i, \text{ and}$$

$$a_i = n_i^{*2} \Big/ n_i.$$

Furthermore,

$$S^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(r_{ij} - \overline{r_i}\right)^2 \Big/ \sum_{i=1}^{K}\left(n_i - 1\right).$$

Note that S^2 is a jackknife variance estimator introduced in Bieler and Williams (1993). Performance of the above trend statistic was evaluated by Peddada et al. (2005).

The NTP uses a continuity-corrected poly-3 trend test statistic, defined as

$$T_1 = \frac{|A| - CF}{B}, \qquad [2]$$

where

$$CF = \frac{1}{2}\max_{i \geq 2}\left[\frac{n_i^*}{n_i}\left(d_i - \overline{d}\right) - \frac{n_{i-1}^*}{n_{i-1}}\left(d_{i-1} - \overline{d}\right)\right]$$

$$\text{and } \overline{d} = \frac{\sum_{i=1}^{K} a_i d_i}{\sum_{i=1}^{K} a_i}. \qquad [3]$$

The null hypothesis of no chemical effect on tumor incidence is rejected in favor of the alternative that there is a positive trend in dose if $A > 0$ and T_1 exceeds the $(1- \alpha)$th percentile of a standard normal distribution.

As an alternative to the poly-3 survival-adjusted CA trend test statistic, Peddada et al. (2005) introduced the following trend test statistic:

$$T_{ISO} = \frac{\tilde{\pi}_k - \tilde{\pi}_1}{S\sqrt{\frac{n_1}{n_1^{*2}} + \frac{n_K}{n_K^{*2}}}}, \qquad [4]$$

where

$$\tilde{\pi}_i = \max_{s \leq i}\min_{t \geq i}\frac{\sum_{j=s}^{t} n_j^* \hat{\pi}_j}{\sum_{j=s}^{t} n_j^*} \qquad [5]$$

are the isotonized values of $\hat{\pi}_i$ under the order restriction that π_i values are nondecreasing with dose. This statistic performs well in terms of power when π_i values are monotonically, but nonlinearly, increasing with dose, whereas the poly-3 survival-adjusted CA trend test performs well when π_i values are linearly increasing with dose (Peddada et al. 2005).

Motivated by these observations, we propose a hybrid statistic that draws on both of these procedures so that the resulting test statistic has improved power in all situations; that is, the proposed statistic is more likely to detect a dose-related trend, if it exists, than is the poly-3 survival-adjusted CA trend test. First, we note that in some instances it is possible for mortality rates in one or more dose groups to be substantially higher than in the control group. In such cases,

$$\sqrt{\left(n_1 / n_1^{*2}\right) + \left(n_K / n_K^{*2}\right)}$$

in the denominator of the above test statistic may inflate the false-positive or type I error rates. For this reason, we modify the trend statistic T_{ISO} of Peddada et al. (2005) as

$$T_2 = \frac{\tilde{\pi}_K - \tilde{\pi}_1}{S\sqrt{2\max\left(n_1 / n_1^{*2},\ n_K / n_K^{*2}\right)}} \qquad [6]$$

and propose the maximum of the NTP trend test statistic and the modified isotonic regression-based trend statistic,

$$T = \max\left(T_1,\ T_2\right), \qquad [7]$$

as the test statistic for testing a dose-related nondecreasing trend in tumor incidence rate. Note that T is the larger of T_1, a test statistic for the linear trend in survival-adjusted proportions, and larger than T_2, a test statistic for the largest difference in survival-adjusted proportions after standardization to a nondecreasing pattern.

We approximate the distribution of T under the null hypothesis that there is no difference in tumor incidence rates among the dose groups as follows. We generate K independent standard normal random deviates, $X_1, X_2, . . . X_K$, and compute

$$V_1 = \sum_{i=1}^{K}\frac{\left(d_i - \overline{d}\right)X_i}{\sqrt{\sum_{i=1}^{K}\left(d_i - \overline{d}\right)^2}},$$

$$\hat{X}_i = \max_{s \leq i}\min_{t \geq i}\frac{\sum_{j=s}^{t} X_j}{t - s + 1}, \text{ and}$$

$$V_2 = \frac{\hat{X}_K - \hat{X}_1}{\sqrt{2}}. \qquad [8]$$

In the above expression, $\hat{X}_i$, $i = 1, 2, . . . , K$, are the "isotonized" values of X_i We approximate the null distribution of T by simulating the distribution of $\max(V_1, V_2)$ from which p-values or critical values for the trend test can be obtained.

Although all our computations were performed in FORTRAN language, we are in the process of developing a JAVA-based software for implementing the methodology introduced in this article. For additional details regarding the software, please contact the authors.

Multiple pairwise comparisons with the control group. In addition to performing a trend test, scientists at the NTP are often interested in performing pairwise comparisons between each of the dose groups and the control group. In this situation, the null hypothesis is that the tumor incidence rates in the dose groups are no larger than the rate in the control group; the alternative hypothesis is that the tumor incidence rate in at least one dose group is strictly larger than in the control group. In order-restricted inference terminology, this alternative hypothesis is known as simple tree order. Currently, the NTP performs pairwise comparisons of each dose group with the control group by applying test statistic T_1 on two groups, the control group and the particular dose group of interest. Although comparisons are generated between the control group and each of the dose groups, NTP researchers are particularly interested in whether the tumor incidence rate in either of the top two dose groups exceeds the rate in the control group. This approach does not account for multiple comparisons between each of the dose groups and the control group. In the following, we propose a simple test statistic derived from the order-restricted inference methodology introduced by Hwang and Peddada (1994) and used by Peddada et al. (2001).

Using the procedure described by Hwang and Peddada (1994), lifetime tumor rates for the dose groups are estimated by $\tilde{\tilde{\pi}}_1 = \tilde{\pi}_1$ for the control group, and $\tilde{\tilde{\pi}}_1 = \max(\tilde{\pi}_1, \hat{\pi}_1)$, $i \geq 2$.

To illustrate the above formula, suppose the poly-3 tumor rates for the four test groups are $\hat{\pi}_1 = 0.3$, $\hat{\pi}_2 = 0.2$, $\hat{\pi}_3 = 0.4$, $\hat{\pi}_4 = 0.35$, with corresponding poly-3 adjusted sample sizes of $n_1^* = 44.2$, $n_2^* = 45.6$, $n_3^* = 40$, $n_4^* = 41$, respectively. Then, using the formula (Equation 5) for $\tilde{\pi}_1$ we have

$$\begin{aligned}\pi_1 = \min\Bigg\{0.3, &\frac{44.2\times0.3+45.6\times0.2}{44.2+45.6},\\ &\frac{44.2\times0.3+45.6\times0.2+40\times0.4}{44.2+45.6+40},\\ &\frac{44.2\times0.3+45.6\times0.2+40\times0.4+41\times0.35}{44.2+45.6+40+41}\Bigg\}\\ = 0.249.\end{aligned} \quad [9]$$

Accordingly, $\tilde{\tilde{\pi}}_2 = \max(0.249, 0.2) = 0.249$, $\tilde{\tilde{\pi}}_3 = \max(0.249, 0.4) = 0.4$, $\tilde{\tilde{\pi}}_2 = \max(0.249, 0.35) = 0.35$.

To derive the test statistic and its distribution under the null hypothesis that all dose groups have the same tumor rates, we first generate K independent standard normal random deviates $X_1, X_2, \ldots, X_K$. Let

$$\begin{aligned}\tilde{X}_1 &= \min_{1\leq i\leq K}\sum_{j=1}^{i}\frac{X_j}{i},\\ \tilde{X}_i &= \max\left(\tilde{X}_i, X_i\right), i \geq 2,\\ V_3 &= \max_{i\geq 2}\frac{\tilde{X}_i - \tilde{X}_1}{\sqrt{2}}.\end{aligned} \quad [10]$$

The proposed test statistic for comparing the ith dose group with the control group is

$$R_i = \frac{\tilde{\pi}_i - \tilde{\pi}_1}{S\sqrt{2\max\left(n_1/n_1^{*2}, n_i/n_i^{*2}\right)}}. \quad [11]$$

Approximate critical values for R_2, R_3, . . ., R_K can be obtained from the simulated distribution of V_3. The proposed pairwise procedure rejects the null hypothesis if $R = \max(R_2, R_3, \ldots, R_K)$ exceeds the critical value derived from the distribution of V_3. This is a nonparametric analogue of Tukey's honestly significant differences post hoc multiple comparisons procedure commonly applied in analysis of variance settings.

Results

Simulation study. We conducted an extensive simulation study to compare the performance of the proposed trend test, T, with the NTP trend test. A total of 750 nonnull configurations and 150 null configurations, similar to those commonly encountered in the NTP rodent bioassays, were simulated. All simulation results reported in this article are based on 10,000 simulation runs, and the nominal level of significance is $\alpha = 0.05$.

Simulation parameters were patterned after the NTP rodent cancer bioassays. We considered a total of three dose groups (low, medium, and high) and a control group, with 50 animals assigned to each group. As described by other authors (e.g., Dinse 1991; Peddada et al. 2001, 2005), for each animal in the ith dose group, $i = 1, 2, 3, 4$, we generated realizations of two independent Weibull random variables, Y_{i1} and Y_{i2}, where Y_{i1} represented the time to tumor onset and Y_{i2} represented the time to death from natural causes. The survival function of Y_{ij}, $i = 1, 2, 3, 4$ and $j = 1, 2$ is given by $P(Y_{ij} > t|\mathrm{d}_i) = \exp(-\psi_j\phi_{ij}^{d_i}t^{\gamma_j})$. We simulated these random variables such that the duration of the study was 24 months, which is typical of the NTP rodent bioassays. We simulated two dose patterns, 2-fold dose spacing and 5-fold dose spacing, namely, $(d_1, d_2, d_3, d_4) = (0, 0.5, 1, 2)$ and $(0, 0.1, 0.5, 2.5)$.

As previously described (Peddada et al. 2005), we considered constant dose effect on mortality; that is, $\phi_{i2} = \phi_2$, with patterns of $\phi_2 = 1$ (no effect), 1.5, 2, 2.5, and 3 (severe effect). We set the mortality shape parameter at $\gamma_2 = 5$ and baseline mortality scale parameter at $\psi_2 = 4.479 \times 10^{-8}$ so that 70% of the animals in the control group survived to the end of the 2-year study, a rate often observed.

The three tumor onset shape parameter (γ_1) values considered in this study were 1.5, 3, and 6. Poly-3 survival adjustments are based on the assumption that the true tumor onset is Weibull with shape parameter $\gamma_1 = 3$ (Portier and Bailer 1989). Thus, the ideal situation for the poly-3 survival correction is $\gamma_1 = 3$. We considered five different background tumor rates, π_1, ranging from rare (0.001, 0.01, 0.05) to common (0.15, 0.30). Values of the baseline tumor onset scale parameter, ψ_1, corresponding to each π_1 are given in Table 1. Finally, we chose six different sets of the effect of dose on tumor onset, ϕ_{i1}, for each of the five background tumor rates; values of ϕ_{i1} are given in Table 2. In each case, the null hypothesis corresponds to the case when the incidence rates are all equal; that is, the ratios are (1:1:1:1). Thus, a total of 375 nonnull and 75 null configurations were considered for each of the two dose spacings.

Results of the simulation study are represented by scatter plots of false-positive error rates (or power) with the NTP procedure on the horizontal axis and the proposed procedure on the vertical axis. For the trend tests, false-positive error rates are summarized in Figure 1

Table 1. Patterns of tumor onset shape parameter (γ_1), and tumor onset scale parameter (ψ_1) by background tumor rate (π_1).

	Background tumor rates (π_1)				
Variables	0.001	0.01	0.05	0.15	0.30
$\gamma_1 = 1.5$					
ψ_1	9×10^{-6}	9×10^{-5}	4.7×10^{-4}	15×10^{-4}	33×10^{-4}
$\gamma_1 = 3$					
ψ_1	8×10^{-8}	8×10^{-7}	4.2×10^{-6}	13.4×10^{-6}	29.7×10^{-6}
$\gamma_1 = 6$					
ψ_1	6.5×10^{-12}	6.5×10^{-11}	32.5×10^{-11}	10.4×10^{-10}	23.2×10^{-10}

Table 2. Patterns of tumor incidence ratios (ϕ_{11}:ϕ_{21}:ϕ_{31}:ϕ_{41}) for the four dose groups by background tumor rate (π_1).

	Tumor incidence ratio (ϕ_{11}:ϕ_{21}:ϕ_{31}:ϕ_{41})		
Dose–effect set	Very rare tumors (π_1 = 0.001, 0.01)	Somewhat rare tumors (π_1 = 0.05)	Common tumors (π_1 = 0.15, 0.30)
1	1:1:1:1	1:1:1:1	1:1:1:1
2	1:1:1:10	1:1:1:4	1:1:1:2
3	1:1:10:10	1:1:4:4	1:1:2:2
4	1:10:10:10	1:4:4:4	1:2:2:2
5	1:5:5:10	1:1.5:1.5:4	1:1.5:1.5:2
6	1:5:10:15	1:2:3:4	1:1.25:1.75:2

and powers in Figure 2. For the pairwise comparison procedure, false-positive error rates are summarized in Figure 3. In each case, the diagonal line represents the line of equality between the two tests. The horizontal and vertical lines in Figures 1 and 3 are drawn at a distance of $0.05 + 1.645\sqrt{(0.05 \times 0.95)/10{,}000}$ from the origin. In Figures 1 and 3, points falling to the right of the vertical line indicate instances in which the NTP procedure exceeds the nominal level of 0.05, and points falling above the horizontal line correspond to instances in which the proposed test exceeds the nominal level of 0.05. In Figure 2, points falling below and to the right of the diagonal line correspond to instances in which the NTP trend test has more power than the proposed trend test, whereas points falling above and to the left of the diagonal line correspond to instances in which the proposed trend test has more power than the NTP trend test. To reduce clutter in the plots, we tested equality of the false-positive error rates (or power) of the NTP procedure and the proposed procedure using a two-sample *z*-test for proportions, and we plotted only those points for which there was a significant difference between the NTP test and the proposed test at the 5% level of significance.

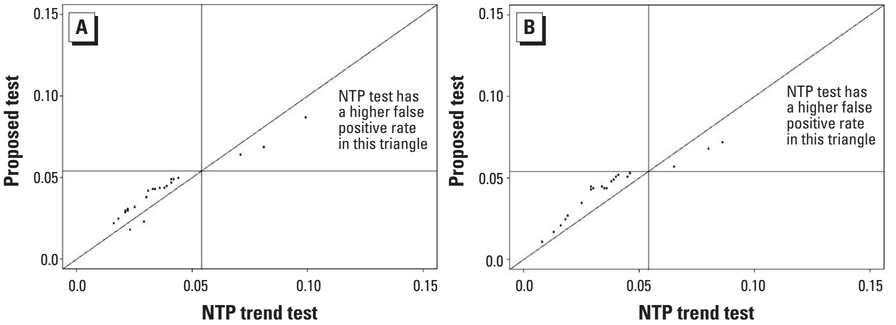

Figure 1. Comparison of false-positive error rates for 2-fold dose spacing (*A*) and 5-fold dose spacing (*B*). The false-positive error rate of the proposed trend test was plotted against the false-positive error rate of the NTP trend test. We plotted only the cases where the two procedures differed significantly in terms of false-positive error rates; results are based on 10,000 simulation runs per configuration. The plot suggests that the false-positive rate of the proposed test is always less than that of the NTP trend test when both tests exceed the nominal level.

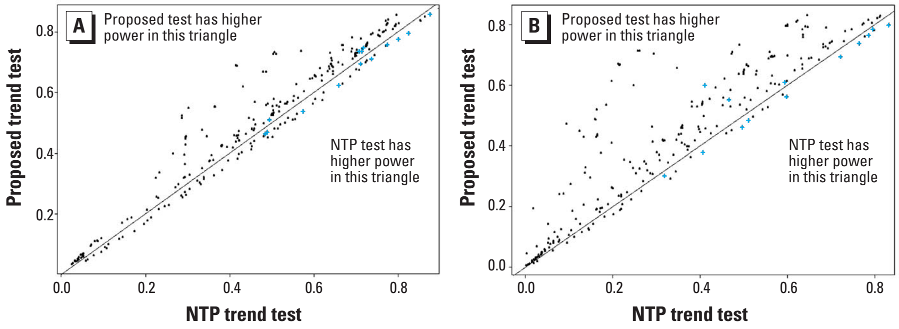

Figure 2. Comparison of power for 2-fold dose spacing (*A*) and 5-fold dose spacing (*B*). In each panel the power of the proposed trend test was plotted against the power of the NTP trend test. Plus symbols (+) indicate cases in which the false-positive error rate of the NTP trend test exceeds both the false-positive error rate of the proposed trend test and 0.05. We plotted only the cases where the two procedures differed significantly in terms of power; results are based on 10,000 simulation runs per configuration. In both (*A*) and (*B*), the "bow"-shaped pattern pointing in the northwest direction, with very few points below the diagonal line, suggests that the proposed test has generally higher sensitivity to detect real trends than the NTP trend test. The gains are substantial as the dose spacing increases from 2- to 5-fold.

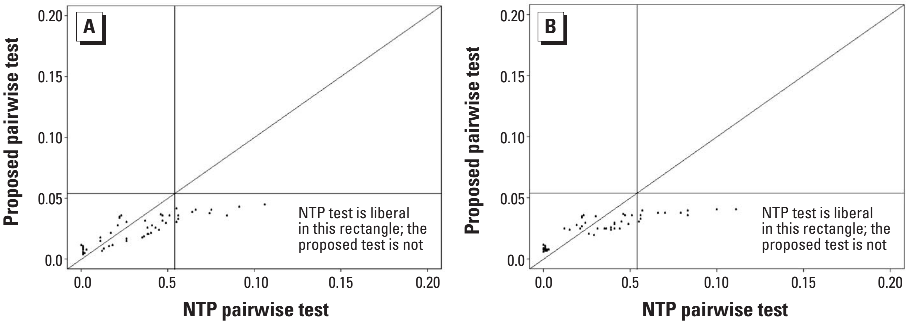

Figure 3. Comparison of false-positive error rates for pairwise tests for 2-fold dose spacing (*A*) and 5-fold dose spacing (*B*). The false-positive error rate of the proposed pairwise comparisons procedure was plotted against the false-positive error rate of the NTP pairwise comparisons procedure for comparing the two highest dose groups against the control. We plotted only the cases where the two procedures differed significantly in terms of false-positive error rates; results are based on 10,000 simulation runs per configuration. In both (*A*) and (*B*), all points are below the horizontal line at 0.054. This suggests that the false-positive rate of the proposed procedure is almost always less than that of NTP pairwise procedure and never exceeds the nominal 0.05 significance level.

For the 75 null patterns considered in this simulation study, there were 23 patterns where the two tests had significantly different false-positive error rates (Figure 1). This result was observed for 2-fold spacing as well as 5-fold spacing. The proposed test was rarely more liberal than the NTP trend test when both tests exceeded the nominal level; that is, the false-positive rate of the proposed test never exceeded that of the NTP trend test. Furthermore, the NTP trend test was more liberal than the proposed test for common tumors ($\pi_1 \geq 0.15$) considered in this study. Although we only plotted the cases for which the false-positive error rates of the two tests differed significantly, the false-positive error rate of the proposed trend test never exceeded 0.087, and that of the NTP trend test never exceeded 0.099.

The power of the two tests differed significantly in 270 of the 375 nonnull dose patterns for 2-fold spacing (Figure 2A). In approximately 70% of these 270 patterns, the proposed trend test had higher power than did the NTP trend test. Thus, a large number of points in Figure 2A are above the diagonal line. Further, in 15 of the 270 patterns (about 6%), the false-positive error rate of the NTP trend test exceeded the nominal 0.05 significance level and was significantly higher than that of the proposed test. These cases are denoted by a "+" in Figure 2A. The gain in power for the proposed test was as high as 0.275 (0.69 for the proposed test vs. 0.415 for the the NTP trend test), a relative gain of 66%. In contrast, the best gain observed for the the NTP trend test was 0.048 (0.502 for the the NTP trend test vs. 0.454 for the proposed test), a modest relative gain of < 10%.

The power gains made by the proposed test were even more substantial for 5-fold dose spacing (Figure 2B). Power of the two tests differed significantly in 264 of the 375 nonnull patterns, and the proposed test had higher power in almost 85% of these patterns. Thus, most points in Figure 2B are above the diagonal.

Further, as we observed with the 2-fold dose spacing, in 13 of the 264 patterns (about 5%) the false-positive error rate of the the NTP trend test exceeded the nominal 0.05 significance level and was significantly higher than that of the proposed test. As in Figure 2A, these cases are denoted by a "+." The gain in power for the proposed test was as high as 0.460 (0.671 for the proposed test vs. 0.211 for the NTP trend test), > 300%. In contrast, the best gain observed for the NTP trend test was 0.038 (0.331 for the NTP trend test vs. 0.293 for the proposed test), a modest relative gain of < 12%.

In cases for which tumor incidence rates increased monotonically, but not linearly, with dose, the proposed trend test performed better than the NTP trend test in terms of both power and false-positive error rate. As expected, the NTP trend test performed better than the proposed test in cases for which tumor incidence rates increased linearly with dose. But even in such cases, the gains made by the NTP trend test were modest. Furthermore, the false-positive error rate of the NTP trend test often exceeded the nominal 0.05 significance level.

For the null configurations described above, we also compared false-positive error rates of the proposed pairwise comparisons procedure with the NTP procedure for pairwise comparisons between the medium- and high-dose groups with the control group. Figure 3 shows that the proposed method maintained false-positive error rates at or below the nominal 0.05 level, whereas the NTP procedure was often liberal, exceeding the nominal level of 0.05. Although we plotted only the cases in which the false-positive error rates of the two tests differed significantly, the proposed pairwise test never exceeded 0.05, whereas the NTP pairwise test had false-positive error rates as high as 0.11.

An NTP example. As part of an NTP bioassay on isoprene, female F344/N rats were exposed to isoprene for 2 years through inhalation (NTP 1999). Isoprene is a naturally occurring compound in plants, as well as a by-product of ethylene production. It is similar in structure to 1,3-butadiene, a potent rodent carcinogen.

Fifty female rats were exposed to 0, 220, 700, or 7,000 ppm isoprene; 19, 35, 32, and 32, respectively, developed mammary gland fibroadenomas. Survival-adjusted tumor proportions showed a plateau-shaped response, with 44%, 74%, 74%, and 73%, respectively, of the animals developing fibroadenomas. The NTP trend test gave a p-value of 0.105, whereas each dosed group differed from the control group at $p < 0.002$. Because of the wide dose spacing and the plateau-shaped response beginning at the low dose of 220 ppm, the NTP trend test was not sensitive enough to detect the dose-related response.

The proposed trend test provided a significant dose-related trend in mammary gland fibroadenomas with a p-value of 0.0014. As indicated in our simulation study discussed above, this statistic is capable of detecting monotonic nonlinear trends with dose and is not affected by wide dose spacing. Furthermore, using the proposed method for pairwise comparisons, each dose group differs from the control group at $p < 0.005$. From our simulations, we can be confident that, among all of the pairwise comparisons with the control group, the overall false-positive rate of 0.05 is not exceeded.

Discussion

We have presented a trend test for tumor incidence data that takes advantage of the strengths of the CA trend test when the dose–response relationship is linear and also takes advantage of strengths of nonparametric order-restricted methods when the dose–response relationship is monotonic but nonlinear. Most important, the false-positive rate of the proposed test rarely exceeds that of the NTP trend test when both tests exceed the nominal level, yet in many instances the proposed test outperforms the NTP trend test in terms of power. Further, we have also provided a simple procedure for performing pairwise comparisons between each of the dose groups and the control group;, this procedure controls the overall false-positive error rates when conducting multiple tests.

Because NTP rodent bioassay data are used by federal and state agencies to assist in formulating regulatory policies, it is crucial for the statistical methods to be powerful enough to detect dose-related trends when they exist. Equally important, these methods should not produce excessive false-positive findings. The trend test and the multiple comparisons procedure that we have proposed here make important steps in both directions. When dose–response relationships are monotonic but nonlinear, the proposed trend test is more powerful than is the NTP trend test. Although both trend tests can exceed the nominal 0.05 level under some circumstances, the false-positive error rate of the proposed trend test is almost always less than that of the NTP trend test when both tests exceed the nominal level. Therefore, the occurrences of false positives will be reduced with use of the proposed trend test. Furthermore, the NTP pairwise comparisons method does not correct for multiple comparisons and often exceeds the nominal 0.05 level in pairwise comparisons of each of the dose groups to the control group, particularly for common tumors. The proposed pairwise comparisons method controls the false-positive rate so that it stays near (or less than) 0.05 under a wide range of situations that we commonly encounter in NTP studies. Thus, these proposed methods should provide more accurate decisions about the potential carcinogenic effects of chemicals.

References

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Bailer A, Portier C. 1988. Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics 44:417–431.

Bieler G, Williams R. 1993. Ratio estimates, the delta method, and quantal response tests for increased carcinogenicity. Biometrics 49:793–801.

Cochran W. 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10:417–451.

Dinse G. 1991. Constant risk differences in the analysis of animal tumorigenicity data. Biometrics 47:681–700.

Hwang J, Peddada S. 1994. Confidence interval estimation subject to order restrictions. Ann Stat 22:67–93.

NTP. 1999. Toxicology and Carcinogenesis Studies of Isoprene (CAS No. 78-79-5) in F344/N Rats (Inhalation Studies). Technical Report No. 486. Research Triangle Park, NC:National Toxicology Program.

Peddada S, Dinse G, Haseman J. 2005. A survival-adjusted quantal response test for comparing tumor incidence rates. Appl Stat 54:51–61.

Peddada S, Prescott K, Conaway M. 2001. Tests for order restrictions in binary data. Biometrics 57:1219–1227.

Portier CJ, Bailer AJ. 1989. Testing for increased carcinogenicity using a survival-adjusted quantal response test. Fundam Appl Toxicol 12:731–737.